Can Deleterious Mutations Explain the Time Dependency of Molecular Rate Estimates?

Michael Woodhams
Allan Wilson Centre for Molecular Ecology and Evolution, Massey University, New Zealand

It has recently been observed by Ho et al. (Ho SYW, Phillips MJ, Cooper A, Drummond AJ. 2005. Time dependency of molecular rate estimates and systematic overestimation of recent divergence times. Mol Biol Evol. 22(7):1561–1568) that apparent rates of molecular evolution increase when measured over short timespans. I investigate whether the data are explainable purely by deleterious mutations. I derive an empirical approximation for the persistence of these mutations in a randomly mating population and, hence, derive lower limits on effective population sizes. These limits are high and get higher if additional reasonable assumptions are made. This casts doubt on whether deleterious mutations are able to explain the apparent rate acceleration.

The molecular clock has a long, fruitful, and sometimes controversial history (Bromham and Penny 2003). Typically, a phylogenetic tree is constructed, the clock is calibrated via one or more internal nodes whose dates are fixed by paleontology, and the dates of remaining internal nodes are estimated.

Ho et al. (2005) have observed an acceleration of molecular evolution on short timescales (fig. 1). The rate against time graph resembles the letter “J” fallen on its side, hence has been named the “J-shaped curve” (Penny 2005, Fig. 1) or “the lazy J.” When evolutionary rates follow such a curve, using a molecular clock calibrated on the long-term rate will overestimate the length of recent time intervals. For example, accounting for the J-shaped curve decreases the estimated date of Mitochondrial Eve from 135 to 176 thousand years ago (kya) to 76 kya (Ho et al. 2005).

Discarding the unlikely proposition that mutation rates have been greatly increasing over the last million years or so, we are left needing an explanation of how mutations can be maintained over short periods of time, yet eliminated on long periods. Slightly deleterious mutations are an obvious candidate—over short timescales, they behave like neutral mutations, with prevalence dominated by genetic drift. Over long timescales, their deleteriousness greatly reduces their chance of becoming fixed, so they contribute little to the (long-term) substitution rate (Kimura 1983, eq. 3.11). (My definition of S differs from Kimura’s by a factor of 2, as I am considering a haploid population.) I define $G(S, \tau)$ to be the solution to the diffusion equation, with scaling given by:

$$G(S, \tau) = G(2Nt, t/N) = \lim_{N \to \infty} NF(N, s, t). \quad (1)$$

Consider how many mutations a randomly selected individual has accumulated compared with its ancestor of $t_0$ generations ago. Let $\mu$ be the mean mutation rate (mutations/individual/generation) and assume all mutations have selective advantage $s$. Between times $t$ and $t + dt$, we expect $\mu N \, dt$ new mutations to arise in the population, each of which has probability $F(N, s, t)$ of being present in the descendant. We integrate over time to find the expected number of differences between ancestor and descendant and divide by $t_0$ to find the estimated mutation rate $R$:

$$R(S, \tau_0, \mu) = \lim_{N \to \infty} \frac{1}{t_0} \int_{0}^{t_0} F(N, s, t) \mu N \, dt$$

$$= \frac{1}{t_0} \int_{0}^{t_0} G(S, \tau) \, d\tau. \quad (2)$$

Note that for neutral ($s = 0$) evolution, $R = \mu$ as $F(N, 0, t) = 1/N$.

Now, we need to consider the distribution of the advantage parameter $S$. I take a model where all mutations are deleterious, with $S$ following an exponential distribution with mean $-\psi$ (i.e., the probability density function of $S$ is $e^{-\psi}/\psi$). Define function $R^*$ to be the observed rate function integrated over the distribution of $S$:

$$R^*(\tau, \psi, \mu) = \int_{-\infty}^{0} R(S, \tau, \mu) e^{\psi S} \, dS. \quad (3)$$

Examples of this function are plotted in figure 2c and d. $F(N, s, t)$ can be calculated exactly for small $N$ by modeling the number of mutant alleles in the population as a discrete Markov process. The Markov matrix $M$ corresponding to a single discrete generation is readily calculable. The eigendecomposition of this matrix can be used to express $F$ as a constant plus a sum of decaying exponentials (see Supplementary Material online). (There is an analytic solution for the diffusion approximation [Crow and Kimura 1970, Section 8.6], but it is not useful here as the solution is a series that converges poorly for large $|S|$.)

Key words: deleterious mutation, nearly neutral evolution, mutation rate, substitution rate, rate calibration.

E-mail: m.d.woodhams@massey.ac.nz.

doi:10.1093/molbev/msi077
Advance Access publication September 7, 2006

© The Author 2006. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. All rights reserved.
For permissions, please e-mail: journals.permissions@oxfordjournals.org
I use the exact solutions to $F$ for $-100 \leq S \leq 10$ and $N \leq 500$ to find an approximation for $G$. The approximation is a constant plus up to 3 decaying exponentials. For $S > -30$, the decay parameters are found by interpolation from the exact results for $N = 500$ (from eigendecomposition of the Markov matrix). For $S < -30$, the approximation is

$$G(S, \tau) \approx \left(1 - 4.344|S|^{-1.164}\right) e^{(1.584 + 0.51985)\tau} + \max(0, 0.463 + 0.000264S) e^{(0.589 + 1.0455)\tau}. \quad (4)$$

Now we can compare the predictions of the $R^*$ curve against the J-shaped curve graphs of Ho et al. (2005). Roughly speaking, the ratio between the y intercept of the curve (zero-time mutation rate) and the asymptote (long-term mutation rate) sets $\psi$, and the decay rate sets $N$.

Each data point from Ho et al. (2005) consists of a calibration time, mean rate, and 95% confidence limits on the rate. I have modeled the distribution of the rate estimate as a log normal distribution with mean equal to the given mean and 95% of the weight within the given confidence limits. This model allows me to calculate a likelihood for a given rate curve (parameterize by $N_g$, $\psi$, and $\nu$, where $g$ is the mean generation time and $\nu$ is the asymptotic mutation rate, i.e., the substitution rate, as opposed to $\mu$, the instantaneous mutation rate).

To simplify calculations, I have truncated the range of $S$ to $-1,000 < S < 0$. (Note that we are not considering highly deleterious or lethal mutations. Even at $S = -1,000$ and $N = 10,000$, we have $s = -0.05$, just a 5% reduction in fitness.) The maximum likelihood fits of this model to the Ho et al. (2005) data are shown in table 1.

Notice that very high values of $\psi$ are obtained (most mutations have large negative $S$), and effective populations are on the order of $10^8$ for birds and $10^5$ for primates. Figure 1 shows one of these maximum likelihood curves.

To estimate a rough lower limit for $N_g$, I set trial values of $N_g$ and optimize on the other 2 variables. The point at which the log likelihood is less than the maximum likelihood by 2 is the point at which allowing $N_g$ to vary significantly outperforms the fixed value (likelihood ratio test). This gives limits $N_g \geq 6.3 \times 10^4$ (avian), $N_g \geq 5.0 \times 10^3$ (primate protein), and $N_g \geq 4.9 \times 10^3$ (primate d-loop). Hence, if the J-shaped curve is due to deleterious mutation, we can set approximate effective population limits of $N > 1.6$ million for bird species (assuming $g = 4$ years) and $N \geq 50,000$ for primates (assuming $g = 10$ years). These bounds are effective population sizes—census populations are typically 4–20 times larger (Frankham 1995). (The approximate generation times are justified in the Supplementary Material online.)

The essence of these bounds is this: to explain (by deleterious mutation) the contrast between short-term and long-term rates, it is necessary to have many significantly deleterious mutations. (Note that for a population of 1 million $S = -1,000$ implies $s = -5 \times 10^{-4}$, so these are still minor mutations from an individual’s point of view.) The more deleterious the mutations, the quicker $G(S, \tau)$ decays to its asymptotic value (fig. 2a). The quicker the decay (in scaled time $\tau$), the larger the $N$ must be to match the observed decay due to the timescaling equation $\tau = N t$.

Note that these limits are conservative as they are based on the assumption that there are no strictly neutral mutations. For neutral mutations, $F$, $G$, and $R$ are constant. If some fraction of the asymptotic rate is due to neutral mutations, the remainder (to be explained by deleterious mutations) has a larger contrast between short- and long-term rates, hence requires larger $N$. For example, the 95% confidence limit for the avian data is $N_g \geq 1.1 \times 10^3$, if half of the fixation rate is due to strictly neutral mutations—an increase of 75%. If we are unwilling to accept high ratios of short- to long-term mutation rates and fix the ratio at 100:1 (i.e., set $\psi = 163$), then the avian data gives 95% confidence limit $N_g \geq 2.2 \times 10^3$.

I have not examined the effects of population or generation time varying over time or between lineages. The general rule is that the long-term effective population size is the harmonic mean of the short-term effective population sizes (Wright 1938), so bottlenecks have a disproportionate effect. These results should be taken as indicative only—large uncertainties remain, notably that the error bars in Ho et al. (2005) are large, generation times have not been accurately estimated, a simplistic model of the distribution of $S$ was used, and the effects of variable population sizes not accounted for.

Recently Bazin et al. (2006) have concluded that mitochondrial evolution is dominated by positive selection. If confirmed, this result invalidates my model and rules out deleterious mutations as an explanation for the J-shaped curve.

In conclusion, it appears that large effective populations are required to explain the J-shaped curve purely by deleterious mutations alone. The increase in observed rates in the short term (the height of the J-shaped curve) requires most mutations to be significantly deleterious and, hence, quickly lost from the population. Large
populations are then required to maintain these mutations for the timescales over which the apparent rate is elevated (the length of the J curve’s hook).

Supplementary Material

A supplementary material elaborating on the mathematics and the derivation of the approximation (eq. 4) is available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This research was inspired by discussions with Mike Hendy and David Penny, who also reviewed the manuscript. Simon Ho provided a copy of the data from Ho et al. (2005). I have had useful suggestions and criticisms from them and from Michael Baake, Alexei Drummond, Robert McLachlan, and Hamish Spencer.

Literature Cited


